

# Quantifying causal influences

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf

March 29, 2012

## Abstract

Common methods of causal inference generate directed acyclic graphs (DAGs) that formalize causal relations between  $n$  variables. Given the joint distribution of all these variables, the DAG contains all information about how intervening on one variable would change the distribution of the other  $n - 1$  variables. It remains, however, a non-trivial question how to *quantify* the causal influence of one variable on another one.

Here we propose a measure for causal strength that refers to direct effects and measure the “strength of an arrow” or a set of arrows. It is based on a hypothetical intervention that modifies the joint distribution by cutting the corresponding edge. The causal strength is then the relative entropy distance between the old and the new distribution.

We discuss other measures of causal strength like the average causal effect, transfer entropy and information flow and describe their limitations. We argue that our measure is also more appropriate for time series than the known ones.

Finally, we discuss conceptual problems in defining the strength of *indirect* effects.

## 1 Introduction

Inferring causal relations is among the most important scientific goals since causality, as opposed to mere statistical dependences, provide the basis for reasonable human decisions. During the past decade, it has become popular to phrase causal relations in directed acyclic graphs (DAGs) [1] with random variables (formalizing statistical quantities after repeated observations) as nodes and causal influences as arrows.

We briefly explain this formal setting. Any system in which there are no two-way interactions (neither direct nor indirect) can be formalized as a DAG. Let  $G$  be such a causal DAG with nodes  $X_1, \dots, X_n$ . To simplify notation, we will mostly assume the  $X_j$  to be discrete.  $P(x_1, \dots, x_n)$  denotes the probability mass function of the joint distribution  $P(X_1, \dots, X_n)$ . By replacing sums with integrals, one can get a straightforward generalization to continuous variables. If  $PA_j$  denotes the set of parent variables of  $X_j$  (i.e., its direct causes) in  $G$ , the joint probability factorizes into

$$P_G(x_1, \dots, x_n) = \prod_{j=1}^n P(x_j | pa_j), \quad (1)$$

where  $pa_j$  denotes the values of  $PA_j$ . This factorization is implied [2] by the Markov condition stating that every node  $X_j$  is conditionally independent of its non-descendants, given its parents. According to the Causal Markov Condition, which we take for granted in this paper, DAGs are only considered as possible *causal* DAGs if they render the joint distribution Markovian [3, 1]. Here and throughout the paper, we have implicitly assumed causal sufficiency, i.e., there are no hidden variables that influence more than one of the  $n$  observed variables. Moreover, by slightly abusing the notion of conditional probabilities,

we assume that  $P(X_j|pa_j)$  is also defined for those  $pa_j$  with  $P(pa_j) = 0$ . This means that we also know how the causal mechanisms act on potential combinations of values of the parents that never occur.

Given this formalism, one may wonder about the motivation for defining causal strength. After all, the DAG together with the joint distribution contains the complete causal information: one can easily compute how the joint distribution changes when an external intervention sets some of the variables to specific values [1]. However, describing causal relations in nature by a DAG always requires to decide how detailed the description should be. Depending on the desired precision, one may want to account for some weak causal links or not. This motivates the need for an objective criterion about which arrows are considered weak.

We first discuss some definitions of causal strength that are either known or just come up as straightforward ideas.

**Average causal effect:** Following [1],  $P(Y|do x)$  denotes the distribution of  $Y$  when  $X$  is set to the value  $x$  (it will be introduced more formally in eq. (5)). Note that it only coincides with the usual conditional distribution  $P(Y|x)$  if the statistical dependence between  $X$  and  $Y$  is due to a direct influence of  $X$  on  $Y$ , with no confounding contribution of some latent common cause. If all  $X_i$  are binary variables, causal strength can then be quantified by the Average Causal Effect [4, 1]

$$ACE(X_i \rightarrow X_j) := P(X_j = 1|do X_i = 1) - P(X_j = 1|do X_i = 0).$$

For real-valued variables  $X_i$  that are affected by a binary variable  $X_j$ , the shift of the mean of  $X_j$  that is caused by switching  $X_i$  from 0 to 1. Formally, one considers the difference [5]

$$\mathbb{E}(X_j|do X_i = 1) - \mathbb{E}(X_j|do X_i = 0).$$

This measure only accounts for the linear aspect of an interaction.

**Analysis of Variance (ANOVA):** Let  $X_i$  be caused by  $X_1, \dots, X_{i-1}$ . Without any assumptions, the variance of  $X_i$  can formally be split into the average of the variances of  $X_i$ , given  $X_k$  and the variance of the expectations of  $X_i$ , given  $X_k$ :

$$\text{Var}(X_i) = \text{Var}(X_i|X_k) + \text{Var}(\mathbb{E}(X_i|X_k)).$$

Within the common scenario of drug testing experiments, for instance, the first term describes the variability of  $X_i$  within a group of equal treatments (i.e. fixed  $x_k$ ), while the second one describes how much the means of  $X_i$  vary between different treatments. It is tempting to say that the latter describes the part of the total variation of  $X_i$  that is *caused by* the variation of  $X_k$ , but this is conceptually wrong for non-linear influences and if there are statistical dependences between  $X_k$  and the other parents of  $X_i$  [6, 5].

For linear structure equations,

$$X_i = \sum_{j < i} \alpha_{ij} X_j + N_i \quad \text{with } N_j \text{ being jointly independent,}$$

with additionally assuming  $X_k$  to be independent of the other parents of  $X_i$ , the second term is given by  $\text{Var}(\alpha_{ik} X_k)$ , which indeed describes the amount by which the variance of  $X_i$  decreases when  $X_k$  is set to a fixed value by intervention. In this sense,

$$r_{ik} := \frac{\text{Var}(\alpha_{ik} X_k)}{\text{Var}(X_i)}$$

is indeed the fraction of the variance of  $X_i$  that is *caused by*  $X_k$ . By rescaling all  $X_j$  such that  $\text{Var}(X_j) = 1$ , we have  $r_{ik} = \alpha_{ik}^2$ . Then, the square of the structure coefficients itself can be seen as a simple measure for causal strength.

**(Conditional) Mutual information:** the information of  $X$  on  $Y$  or vice versa is given by [7]

$$I(X; Y) := \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}.$$

The information of  $X$  on  $Y$  or vice versa if  $Z$  is given is defined by [7]

$$I(X; Y | Z) := \sum_{x,y,z} P(x, y, z) \log \frac{P(x, y | z)}{P(x | z)P(y | z)}. \quad (2)$$

There are situations where these expressions (with  $Z$  describing some background condition) can indeed be interpreted as measuring the strength of the arrow  $X \rightarrow Y$ . An essential part of this paper will be devoted to describing the conditions under which this makes sense and how to replace the expressions with other information-theoretic ones for the other cases.

**Granger causality:** Quantifying causal influence between time series (for instance between  $(X_t)_{t \in \mathbb{Z}}$  and  $(Y_t)_{t \in \mathbb{Z}}$ ) is special because one is interested in quantifying the effect of all  $(X_t)$  on all  $(Y_{t+s})$ . If we represent the causal relations by a DAG where every time instant defines a separate pair of variables, then we ask for the strength of a *set of arrows*. If the time instant  $t$  just describes instances of the same variables  $X, Y$ , we leave the regime of i.i.d. sampling.

Reduction of uncertainty of one variable after knowing the other is also the key idea of several related methods for quantifying causal strength in time series. Granger causality in its original formulation uses reduction of variance [8]. A non-linear information-theoretic extension that is in the same spirit is transfer entropy [9]. The latter is basically a conditional mutual information where each variable  $X, Y, Z$  in (2) is replaced with an appropriate set of variables.

We will discuss conceptual problems with the measures known to us and argue why our measure seems to be a better formalization of causal strength.

## 2 Postulates for causal strength

Let us first discuss the properties we would like a measure of causal strength to have. We present four properties that we consider reasonable.

- P0. **Causal Markov condition and arrows with zero strength:** The joint distributions satisfies the Markov condition also after removing all arrows of zero strength. Formally, this is equivalent to

$$\mathfrak{C}_{X \rightarrow Y} = 0 \implies I(X; Y | PA_Y^X) = 0,$$

where  $PA_Y^X$  denotes the causes of  $Y$  other than  $X$ .

- P1. **Mutual information:** If the true causal DAG is given by  $X \rightarrow Y$ , then

$$\mathfrak{C}_{X \rightarrow Y} = I(X; Y).$$

- P2. **Locality:** The strength of  $X \rightarrow Y$  only depends on (1) how  $Y$  depends on  $X$  and its other parents, and (2) the joint distribution of all parents of  $Y$ . Formally, knowing  $P(Y | PA_Y)$  and  $P(PA_Y)$  is sufficient to compute  $\mathfrak{C}_{X \rightarrow Y}$ . For strictly positive densities  $P(y, pa_Y)$ , this is equivalent to knowing  $P(Y, PA_Y)$ .

- P3. **Quantitative causal Markov condition:** If there is an arrow from  $X$  to  $Y$  then the causal influence of  $X$  on  $Y$  is greater or equal than the conditional mutual information between  $Y$  and  $X$ , given all the other parents of  $Y$ , formally

$$\mathfrak{C}_{X \rightarrow Y} \geq I(X; Y | PA_Y^X).$$

Note that P0 follows from P3 because  $Y \perp\!\!\!\perp X | PA_Y^X$  implies that we can drop the link  $X \rightarrow Y$ . We have started with P0 for didactic reasons. We do not claim that *every* reasonable measure of causal strength should satisfy these postulates, but we now explain why we consider them as natural. To this end, we also show that the implications for simple DAGs make sense.

**P0:** If the purpose of our measure of causal strength is to quantify relevance of an arrow then arrows of zero strength must be irrelevant. In particular, removing such an arrow  $X \rightarrow Y$  does not yield a DAG that is ruled out by the causal Markov condition. Since  $PA_Y^X$  is the set of parents in the simplified DAG  $G'$ , we obtain

$$Y \perp\!\!\!\perp ND_Y | PA_Y^X, \quad (3)$$

where  $ND_Y$  denotes the non-descendants of  $Y$  in  $G'$ . Note that  $X$  cannot be a descendant of  $Y$  in  $G'$  because it has been a parent in the original DAG  $G$ . Thus, (3) implies

$$Y \perp\!\!\!\perp X | PA_Y^X.$$

Hence, we conclude

$$I(Y; X | PA_Y^X) = 0. \quad (4)$$

**P1:** The mutual information actually measures the strength of statistical dependences. Since all these dependences are generated by the influence of  $X$  on  $Y$  (and not by a common cause or  $Y$  influencing  $X$ ), it makes sense to measure causal strength by strength of dependences. Note that mutual information  $I(X; Y) = H(Y) - H(Y|X)$  also quantifies the variability in  $Y$  that is due to the variability in  $X$ , see also §A.3.

*Mutual information versus channel capacity.* Given the premise that causal strength should be an information like quantity, a natural alternative to mutual information is the capacity of the information channel  $x \mapsto P(Y|x)$ , i.e. the maximum over all values of mutual information  $I_{Q(X)}(X; Y)$  for all input distributions  $Q(X)$  of  $X$  when keeping the conditional  $P(Y|X)$ .

While channel  $I(X; Y)$  quantifies the observable dependences, channel capacity quantifies the strength of the strongest dependences that can be generated using the information channel  $P(Y|X)$ . In this sense, that  $I(X; Y)$  quantifies the *factual* causal influence, while channel capacity measures the *potential* influence. Channel capacity also accounts for the impact of setting  $x$  to values that rarely or never occur in the observations. However, this sensitivity regarding effects of rare inputs can certainly be a problem for estimating the effect from sparse data. We therefore prefer mutual information  $I(X; Y)$  as it better assesses to what extent the *frequently observed changes* in  $X$  influence  $Y$ .

**P2:** Locality implies that we can ignore causes of  $X$  when computing  $\mathfrak{C}_{X \rightarrow Y}$ , unless they are at the same time direct causes of  $Y$ . Likewise, other effects of  $Y$  are irrelevant. Moreover, it does not matter *how* the dependences between the parents are generated (which parent influences which one or whether they are effects of a common cause), we only need to know their joint distribution with  $X$ .

Violations of locality would have paradoxical implications. For example, variable  $Z$  should clearly be irrelevant in DAG a) in Figure 1. Otherwise,  $\mathfrak{C}_{X \rightarrow Y}$  would depend on the mechanism that generates the distribution of  $X$ , while we are actually concerned with the information flowing from  $X$  to  $Y$  instead of the one flowing *to*  $X$  from other nodes. Likewise, (see DAGs in Figure 1 b) and c)) it is irrelevant whether  $X$  and  $Y$  have further effects.

**P3:** The postulate quantitatively extends (4): The arrow  $X \rightarrow Y$  is the only reason for the conditional dependence  $I(Y; X | PA_Y^X)$  being non-zero, hence it is natural postulating that its strength cannot be smaller than the dependence that it generates.



Figure 1: DAGs for which the (conditional) mutual information is a reasonable measure of causal strength: For a) to c), our postulates imply  $\mathfrak{C}_{X \rightarrow Y} = I(X; Y)$ . For c) we will obtain  $\mathfrak{C}_{X \rightarrow Y} = I(X; Y | Z)$ .

Note that ACE and ANOVA are already ruled out by P0. Consider a relation between three binary variables  $X, Y, Z$ , where  $Y = X \oplus Z$  with  $X$  and  $Z$  being unbiased and independent. Then changing  $X$  has no influence on the statistics of  $Y$ . Likewise, knowing  $X$  does not reduce the variance of  $Y$ . To satisfy P0, we would need modifications that we do observe an influence of  $X$  on  $Y$  for each fixed value  $z$  rather than marginalizing over  $Z$ . We will not consider ACE and ANOVA any further and now discuss information theoretic approaches.

### 3 Problems of known definitions

Our definition of causal strength is presented in Section 4. This section discusses problems with three measures of causal strength: mutual information, transfer entropy [9] and information flow [10].

#### 3.1 Mutual information and conditional mutual information

It is sufficient to consider a few simple DAGs to illustrate why mutual information and conditional mutual information are *not* suitable measures of causal strength, despite the fact that they arise in special cases (see also P1). Quantifying causal strength in Figure 2ab is already as difficult as the general case because the main challenge in defining  $\mathfrak{C}_{X \rightarrow Y}$  is that dependences between  $X$  and other parents of  $Y$  generate dependences between  $X$  and  $Y$  that interfere with the influence of  $X$  and  $Y$ .

*Mutual information is not suitable in Figure 2a.* It is clear that  $I(X; Y)$  is inappropriate because part of the dependency is due to the common cause  $Z$  and  $I(X; Y) \neq 0$ , which can be arbitrarily large even if the arrow  $X \rightarrow Y$  is missing. Moreover, DAG a) in Figure 2 contains d) in Figure 1 as limiting case (when the arrow  $Z \rightarrow X$  gets weaker), where P3 requires causal strength to majorize the *conditional* information  $I(X; Y | Z)$ . Then causal strength cannot be given by  $I(X; Y)$  because it is easy to construct transition probabilities for DAG d) where  $I(X; Y) = 0$  and  $I(X; Y | Z) \neq 0$ .

*Conditional mutual information is not suitable for Figure 2a.* Showing that  $I(X; Y | Z)$  is, nevertheless, inappropriate for DAG a) in Figure 2 is more subtle – after all it qualitatively behaves correctly in the sense that it vanishes when the arrow  $X \rightarrow Y$  disappears. To see the wrong quantitative behavior, we consider the limiting case where the influence from  $Z \rightarrow Y$  gets weaker until it disappears completely at which point we obtain DAG a) in Figure 1. We have already discussed that our postulates imply  $\mathfrak{C}_{X \rightarrow Y} = I(X; Y)$  and not  $I(X; Y | Z)$  for this case.

Note that for DAG a) in Figure 1,  $I(X; Y)$  is larger than  $I(X; Y | Z)$  because  $Y \perp\!\!\!\perp Z | X$  implies  $I(X; Y) = I(Z; Y) + I(Y; X | Z)$  using standard properties of mutual information



Figure 2: DAGs for which finding a proper definition of  $\mathfrak{C}_{X \rightarrow Y}$  is challenging.

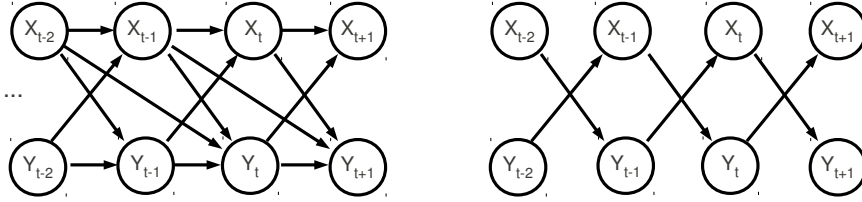


Figure 3: Left: Typical causal DAG for two time series with mutual causal influence. The structure is acyclic because instantaneous influences are excluded. Right: counter example in [10]. Transfer entropy vanishes if all arrows are copy operations although the time series strongly influence each other.

[7]. This shows that  $I(X; Y | Z)$  underestimates the causal strength at least for the limiting case where  $Z \rightarrow Y$  disappears completely. We will later see that  $I(X; Y | Z)$  is also too small for the case where  $Z \rightarrow Y$  is strong enough to be relevant.

*Both conditional and unconditional mutual information are unsuitable for Figure 2b.* The reasons are similar to those for DAG a).  $I(X; Y)$  would measure the overall effect of  $X$  on  $Y$ , which is partly due to the arrow  $X \rightarrow Y$  and partly due to the path over  $Z$ .

Conditional information  $I(X; Y | Z)$  behaves quantitatively correct in the sense that it vanishes if the arrow  $X \rightarrow Y$  disappears. However, it underestimates the strength of the arrow. Consider the limiting case where the arrow  $Z \rightarrow Y$  disappears, yielding DAG c) in Figure 1. In the limit we wish to obtain  $I(X; Y)$ , which is again larger than  $I(X; Y | Z)$  due to  $Y \perp\!\!\!\perp Z | X$  (see above).

### 3.2 Transfer entropy

Transfer entropy [9] is intended to measure the influence of one time-series on another one. Let  $(X_t, Y_t)_{t \in \mathbb{Z}}$  be a bivariate stochastic process where  $X_t$  influence some  $Y_s$  with  $s > t$ , see figure 3, left. Then transfer entropy is defined as the following conditional mutual information:

$$I(X_{(-\infty, t-1]} \rightarrow Y_t | Y_{(-\infty, t-1]}) := I(X_{(-\infty, t-1]}; Y_t | Y_{(-\infty, t-1]}).$$

It measures the amount of information the past of  $X$  provides about the present of  $Y$  given the past of  $Y$ . To quantify causal influence by conditional information relevance is also in the spirit of Granger causality, where information is usually understood in the sense of the amount of reduction of the linear prediction error.

*Transfer entropy is an unsatisfactory measure of causal strength.* [10] have pointed out that this information relevance fails to quantify causal influence for the following toy

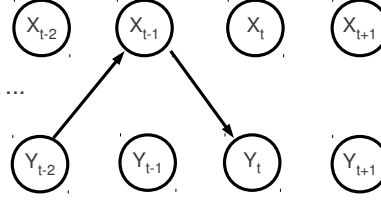


Figure 4: Time series with only two causal arrows, where transfer entropy fails satisfying our postulates.

model: Assume the information from  $X_t$  is perfectly copied to  $Y_{t+1}$  and the information from  $Y_t$  to  $X_{t+1}$ . Then the past of  $Y$  is already sufficient to perfectly predict the present value of  $Y$  and the past of  $X$  does not provide any further information. Therefore, transfer entropy vanishes although both variables heavily influence each other.

*Transfer entropy violates our postulates.* We now show that transfer entropy yields 0 bits of causal influence in a situation where common sense and our postulates require that causal strength is 1 bit. Since our postulates refer to the strength of a *single* arrow while transfer entropy is supposed to measure the strength of all arrows from  $X$  to  $Y$ , we reduce the DAG such that there is only one arrow from  $X$  to  $Y$ , see figure 4. Then,

$$I(X_{(-\infty, t-1]} \rightarrow Y_t | Y_{(-\infty, t-1]}) = I(X_{(-\infty, t-1]}; Y_t | Y_{(-\infty, t-1]}) = I(X_{t-1}; Y_t | Y_{t-2}).$$

The causal structure coincides with the DAG a) in figure 1 by setting  $Y_{t-2} \equiv Z$ ,  $X_{t-1} \equiv X$ , and  $Y_t \equiv Y$ . With these replacements, transfer entropy yields  $I(X; Y | Z) = 0$  bits instead of  $I(X; Y) = 1$  bit, as required by P1.

Further critical discussion of transfer entropy can be found in [11] in the context of cellular automata dynamics.

### 3.3 Information flow

After arguing that transfer entropy does not properly capture the strength of the impact of interventions, [10] proposes to define causal strength using Pearl's *do* calculus [1]. Given a causal directed acyclic graph  $G$ , Pearl computes the joint distribution obtained if variable  $X_j$  is forcibly set to the value  $x_j$  as

$$P(x_1, \dots, x_n | do x'_j) := \prod_{i \neq j} P(x_i | pa_i) \cdot \delta_{x_j, x'_j}. \quad (5)$$

Given three sets of nodes  $X_A$ ,  $X_B$  and  $X_C$  in a directed acyclic graph  $G$ , information flow is computed as

$$I(X_A \rightarrow X_B | do X_C) := \sum_{x_C, x_A, x_B} p(x_C) P(x_A | do x_C) P(x_B | do x_A, do x_C) \log \frac{P(x_B | do x_A, do x_C)}{\sum_{x'_A} P(x'_A | do x_C) P(x_B | do x'_A, do x_C)}$$

To better understand this expression, we first consider the case where the set  $X_C$  is empty. Then we obtain

$$I(X_A \rightarrow X_B) := \sum_{x_A, x_B} P(x_B | do x_A) \log \frac{P(x_B | do x_A)}{\sum_{x'_A} P(x'_A) P(x_B | do x'_A)},$$

which measures the mutual information between  $X_A$  and  $X_B$  obtained when the information channel  $x_A \mapsto P(X_B|do x_A)$  is used with the input distribution  $P(X_A)$ . By using the post-interventional conditional distribution rather than the observed conditional  $x_A \mapsto P(X_B|x_A)$  as information channel, this definition certainly goes the right step from a dependence measure to a measure of causal strength. One may ask why the input  $P(X_A)$  is used for this channel instead of some other distribution of  $X_A$ . Similar questions will arise for our definition of causal strength, too. Here we accept  $P(X_A)$  at least as a straightforward choice, although others could be justified. The question about the appropriate input gets even more important when  $X_C$  is a non-empty set of “background” variables (other causes of  $X_B$ , for instance). It is natural to consider then the information channels  $x_A \mapsto P(X_B|do x_A, do x_C)$  for different  $x_C$  and average the resulting mutual information over all  $x_C$ , weighted by  $P(X_C)$ . [10] decided to choose  $P(X_A|do x_C)$  as input distribution. Although this choice seems to be more in the spirit of describing causality than the choice  $P(X_A|x_C)$  would be, we should mention that it violates our postulate P2, as described below.

*Information flow is an unsatisfactory measure of causal strength.* To quantify  $X \rightarrow Y$  in DAGs a) and b) in Figure 2 by information flow, we may either choose  $I(X \rightarrow Y)$  or  $I(X \rightarrow Y|Z)$ . We show that both choices are inconsistent with our postulates and with our intuitive expectation. We start with  $I(X \rightarrow Y)$  and DAG a). Let  $X, Y, Z$  be binary with  $Y := X \oplus Z$  is the XOR of its causes,  $Z$  be an unbiased coin toss and  $X$  be a faulty copy of  $Z$  with two-sided symmetric error. One easily checks that  $I(X \rightarrow Y)$  is zero in the limit of error probability  $1/2$  (making  $X$  and  $Z$  independent). Nevertheless, dropping the arrow  $X \rightarrow Y$  would violate the Markov condition, in contradiction to P0. For error rate close to  $1/2$ , we still violate P3 because  $I(Y; X|Z)$  is close to 1, while  $I(X \rightarrow Y)$  is close to zero. A similar argument can be constructed for DAG b) in the same figure.

We now consider  $I(X \rightarrow Y|Z)$ . Note that it yields different results for DAGs a) and b) if the joint distribution is the same, in contradiction to P2. This is because  $P(x|do z) = P(x|z)$  for a), while  $P(x|do z) = P(x)$  for b). In other words,  $I(X \rightarrow Y|Z)$  depends on what type of causal relation generated the dependences between the two causes  $X$  and  $Z$ . Apart from being inconsistent with our postulate, we find it unsatisfactory that  $I(X \rightarrow Y|Z)$  tends to zero for the example above if the error rate of copying  $X$  from  $Z$  in DAG a) tends to zero (conditioned on setting  $Z$  to some value, the information passed from  $X$  to  $Y$  is zero because  $X$  attains a fixed value, too). In this limit,  $Y$  is always zero. We argue that the link  $X \rightarrow Y$  still remains important for explaining the behavior of the XOR: without the link, the gate could not always output “zero”, for both values of  $Z$ .

## 4 Defining the strength of causal arrows

### 4.1 Definition in terms of conditional probabilities

This section proposes a way to quantify the causal influence of a set of arrows that yields satisfactory answers in all the cases discussed above. Our measure is motivated by a scenario where the nodes represent different parties communicating with each other via channels. Hence, we think of arrows as physical channels that propagate information between distant points in space, e.g., wires that connect electronic devices. Each such wire connects the output of a device with the input of another one. For the intuitive ideas below, it is also important that the wire connecting  $X_i$  and  $X_j$  physically contains full information about  $X_i$  (which may be more than the information that is required to explain the output behavior  $P(X_j|PA_j)$ ).

We then think of the strength of the arrow  $X_i \rightarrow X_j$  as the impact of corrupting it, i.e., the impact of cutting the wire. To get a well-defined “post-cutting” distribution we have to say what to do with the open end corresponding to  $X_j$ , because it needs to be





Figure 5: Left: deletion of the arrow  $X \rightarrow Y$ . The conditional  $P(Y|X, Z)$  is fed with an independent copy of  $X$ , distributed with  $P(X)$ . The resulting distribution reads  $P_{X \rightarrow Y}(x, y, z) = P(x, z) \sum_{x'} P(y|z, x')P(x')$ . Right: deletion of both incoming arrows. The conditional  $P(Y|X, Z)$  is then fed with the product distribution  $P(X)P(Z)$  instead of the joint  $P(X, Y)$  since the latter would require communication between the open ends. This results in the distribution  $P_{X \rightarrow Y, Z \rightarrow Y}(x, y, z) = \sum_{x', z'} P(x, z)P(y|x', z')P(x')P(z')$ .

fed with some input. It is natural to feed it probabilistically with inputs  $x_i$  according to  $P(X_i)$  because this is the only distribution of  $X_i$  that is locally observable (feeding it with some conditional distribution  $P(X_i|..)$  would assume that the one who cuts the edge would have access to other nodes and not only to the physical state of the channel. Likewise, we define the deletion of a set of arrows by feeding all open ends with the product of the corresponding marginal distributions. Also here we argue that feeding the open ends with some non-product distribution would require communication between the different open ends. Figure 5 visualizes the deletion of one edge (left) and two edges (right).

*Remark 1.* The communication scenario also motivates our choice of mutual information rather than capacity in postulate P1. The capacity of  $X \rightarrow Y$  cannot be computed locally at  $X$ , since it requires that the intervener not only observes  $X$  but also  $Y$ .

We now define the “post-cutting” distribution formally:

**Definition 1** (removing causal arrows). *Let  $G$  be a causal DAG and  $P$  be Markovian with respect to  $G$ . Let  $S \subset G$  be a set of arrows. Set  $PA_j^S$  as the set of those parents  $X_i$  of  $X_j$  for which  $(i, j) \in S$  and  $PA_j^{\bar{S}}$  those for which  $(i, j) \notin S$ . Set*

$$P_S(x_j|pa_j^{\bar{S}}) := \sum_{pa_j^S} P(x_j|pa_j^S, pa_j^{\bar{S}})P_{\Pi}(pa_j^S), \quad (6)$$

where  $P_{\Pi}(pa_j^S)$  denotes for a given  $j$  the product of marginal distributions of all variables in  $PA_j^S$ . Define new joint distribution, the interventional distribution,

$$P_S(x_1, \dots, x_n) := \prod_j P_S(x_j|pa_j^{\bar{S}}). \quad (7)$$

See Figure 5, right, for a simple example with cutting only one edge. Eq. (7) formalizes the fact that each open end of the wires is independently fed with the corresponding marginal distribution, see also Figure 5, right. The modified joint distribution  $P_S$  can be considered as generated by the reduced DAG:

**Lemma 1** (Markovian). *The interventional distribution  $P_S$  is Markovian with respect to the graph  $G_S$  obtained from  $G$  by removing the edges in  $S$ .*

*Proof.* By construction,  $P_S$  factorizes according to  $G_S$  in the sense of (1).  $\square$

**Definition 2** (causal influence of a set of arrows). *The causal influence of the arrows in  $S$  is given by the Kullback-Leibler divergence*

$$\mathfrak{C}_S(P) := D(P\|P_S). \quad (8)$$

If  $S$  is a single edge  $X_k \rightarrow X_l$ , we write  $\mathfrak{C}_{k \rightarrow l}$  instead of  $\mathfrak{C}_{X_k \rightarrow X_l}$ .

Note that  $P_S$  could easily be confused with different distributions that we obtain when the open ends are not fed with the marginal distributions but with conditional distributions. We want to explain this for DAG a) in Figure 2. Define  $\tilde{P}_{X \rightarrow Y}(X, Y, Z)$  by

$$\tilde{P}_{X \rightarrow Y}(x, y, z) := P(x, z)P(y|z) = P(x, z) \sum_{x'} P(y|x')P(x'|z),$$

and recall that replacing  $P(x'|z)$  with  $P(x')$  in the right most expression yields  $P_{X \rightarrow Y}$ . We call  $\tilde{P}_{X \rightarrow Y}$  the “partially observed distribution”. It is the distribution that one erroneously gets when ignoring the influence of  $X$  on  $Y$ :  $\tilde{P}_{X \rightarrow Y}$  is computed according to (1), but uses a DAG where  $X \rightarrow Y$  is missing. The difference between “ignoring” and “cutting” the edge is important for the following reason. By a known rephrasing of mutual information as relative entropy [7] we obtain

$$D(P\|\tilde{P}_{X \rightarrow Y}) = I(X; Y | Z), \quad (9)$$

which, as we have already discussed, is *not* in general a satisfactory measure of causal strength.

## 4.2 Definition via structure equations

Our definition of  $\mathfrak{C}_{i \rightarrow j}$  uses the conditional density  $P(x_j|pa_j)$ . Estimating a conditional density from empirical data requires huge samples or strong assumptions (particularly for continuous variables). Fortunately, however, structure equations (also called functional models [1]) allow for a more direct estimation of causal strength without referring to the conditional distribution.

**Definition 3** (structure equation). *A structure equation is a model that explains the joint distribution  $P(X_1, \dots, X_n)$  by a deterministic dependence*

$$X_j = f_j(PA_j, E_j),$$

where the variables  $E_j$  are jointly independent unobserved noise variables. Note that functions  $f_j$  that correspond to parentless variables can be chosen to be the identity, i.e.,  $X_j = E_j$ .

Suppose that we are given a causal inference method that directly infers the structure equations (e.g., [12, 13]) in the sense that it outputs  $n$ -tuples  $(e_1^j, \dots, e_n^j)$  with  $j = 1, \dots, m$  as well as the functions  $f_j$  from the observed  $n$ -tuples  $(x_1^j, \dots, x_n^j)$ .

**Definition 4** (removing a causal arrow in a structure equation). *Deletion of the arrow  $X_k \rightarrow X_l$  is modeled by (i) introducing an i.i.d. copy  $X'_k$  of  $X_k$  and (ii) subsuming the new random variable  $X'_k$  into the noise term of  $f_l$ . The result is a new set of structure equations:*

$$\begin{aligned} x_j &= f_j(pa_j, e_j) & \text{if } j \neq l, \text{ and} \\ x_l &= f_l(pa_l \setminus \{x_k\}, (x'_k, e_l)), \end{aligned} \quad (10)$$

where we have omitted the superscript  $j$  to simplify notation.

*Remark 2.* For measuring the causal influence of a set of arrows, the procedure works similarly, then we need to introduce jointly independent i.i.d. copies of all variables being at tails of deleted arrows.

*Remark 3.* The change introduced by the deletion only effects  $X_l$  and its descendants, the virtual sample thus keeps all  $x_j$  with  $j < k$ . Moreover, we can ignore all variables  $X_j$  with  $j > l$  due to Lemma 4.

Note that  $x'_k$  must be chosen to be independent of all  $x_j$  with  $j \leq k$ , but, by virtue of the structure equations, not independent of  $x_l$  and its descendants. The new structure equations thus generate  $n$ -tuples of “virtual” observations  $x_1^S, \dots, x_n^S$  from the input

$$(e_1, \dots, (x'_k, e_l), \dots, e_n).$$

We will show below that the  $n$ -tuples generated this way indeed follow the distribution  $P_S(X_1, \dots, X_n)$ . We can therefore estimate causal influence via any method that estimates relative entropy distance using the observed samples  $x_1, \dots, x_n$  and the virtual ones  $\tilde{x}_1, \dots, \tilde{x}_n$ . To illustrate the above scheme, we consider the case where  $Z$  and  $X$  are causes of  $Y$  and we want to delete the edge  $X \rightarrow Y$ . The case where  $Y$  has more than 2 parents follows easily.

**Example 1** (Two parents). *The following table corresponds the observed variables  $X, Z, Y$ , as well as the unobserved noise  $E^Y$  which we assumed to be estimated together with learning the structural equations.*

$$\begin{pmatrix} Z & X & E^Y & Y \\ z_1 & x_1 & e_1^Y & f_Y(z_1, x_1, e_1^Y) \\ z_2 & x_2 & e_2^Y & f_Y(z_2, x_2, e_2^Y) \\ \vdots & & & \vdots \\ z_m & x_m & e_m^Y & f_Y(z_m, x_m, e_m^Y) \end{pmatrix}. \quad (11)$$

To simulate the deletion of  $X \rightarrow Y$  we first generate a list of virtual observations for  $Y$  after generating samples from an i.i.d. copy  $X'$  of  $X$ :

$$\begin{pmatrix} Z & X & X' & E^Y & Y \\ z_1 & x_1 & x'_1 & e_1^Y & f_Y(z_1, x'_1, e_1^Y) \\ \vdots & & & & \vdots \\ z_m & x_m & x'_m & e_m^Y & f_Y(z_m, x'_m, e_m^Y) \end{pmatrix}. \quad (12)$$

A simple method to simulate the i.i.d. copy is to apply some random permutation  $\pi \in S_m$  to  $x_1, \dots, x_n$  and obtain  $x_{\pi(1)}, \dots, x_{\pi(n)}$ , see below.

We then throw away the two noise columns, i.e., the original noise  $E^Y$  and the additional noise  $X'$ :

$$\begin{pmatrix} Z & X & Y \\ z_1 & x_1 & f_Y(z_1, x'_1, e_1^Y) \\ \vdots & & \vdots \\ z_m & x_m & f_Y(z_m, x'_m, e_m^Y) \end{pmatrix}. \quad (13)$$

To see that this triple is indeed sampled from the desired distribution  $P_S(X, Y, Z)$ , we recall that the original structure equation simulates the conditional  $P(Y|X, Z)$ . After inserting  $X'$  we obtain the new conditional  $\sum_{x'} P(Y|x', Z)P(x')$ . Multiplying it with  $P(X, Z)$  yields  $P_S(X, Y, Z)$ , by definition. Using the above samples from  $P_S(X, Y, Z)$  and samples from  $P(X, Y, Z)$  we can estimate

$$\mathfrak{C}_{X \rightarrow Y} = D(P(X, Y, Z) \| P_S(X, Y, Z))$$

using some known schemes for estimating relative entropies from empirical data. It is important (?) that the samples from the two distributions are disjoint, meaning that we need to split the original sample into two halves, one for  $P$  and one for  $P_S$ .

We now show that random permutations simulate an i.i.d. copy  $X'$ . We first observe:

**Lemma 2.** *Let  $X$  be a discrete random variable with probability mass function  $P(x)$ . Given an i.i.d. sample  $x_1, \dots, x_m$ . Let  $\pi \in S_m$  be a random permutation. Then the empirical distribution of  $(x_j, x_{\pi(j)})$  converges for  $m \rightarrow \infty$  weakly to  $P(x)P(x')$ , where  $X'$  is an i.i.d. copy of  $X$ .*

*Proof.* for any functions  $f, g$  the empirical expectations factorize asymptotically, i.e, the probability that

$$\left| \frac{1}{m} \sum_{j=1}^m f(x_j)g(x_{\pi(j)}) - \frac{1}{m} \sum_{j=1}^m f(x_j) \frac{1}{m} \sum_{j=1}^m g(x_{\pi(j)}) \right| \geq \epsilon$$

for a random permutation  $\pi$  converges to zero. Hence, the empirical distribution of  $(x, x')$ -pairs converges to a product measure, which needs to be  $P(x)P(x')$  because we clearly have weak convergence to  $P(x)$  for the marginals.  $\square$

For our application, we need a slightly stronger version that ensures that the permuted sample is also independent of the other parents of  $X_l$ :

**Lemma 3.** *Let  $X, W$  be two random variables with joint density  $P(x, w)$ . Given an iid sample  $(x_j, w_j)$  with  $j = 1, \dots, m$ . Let  $\pi \in S_m$  be a random permutation. Then the empirical distribution of the sample  $(x_j, w_j, x_{\pi(j)})$  converges weakly to  $P(x, w)P(x')$  where  $X'$  is an i.i.d. copy of  $X$ .*

*Proof.* Using vector valued random variables in Lemma 2, we obtain  $P(X, E)P(X', E')$  by jointly permuting  $x, e$ . Then the statement follows by marginalizing over  $E'$ .  $\square$

Lemma 3 shows that we can indeed generate  $X'$  the way proposed above.

### 4.3 Properties of causal strength

This subsection shows that our definition of causal strength satisfies all our postulates. In proving this, we observe at the same time some other useful properties. We don't need to prove P0 because it is implied by P3.

**P1:** One easily checks  $\mathfrak{C}_{X \rightarrow Y} = I(X; Y)$  for the 2-node DAG  $X \rightarrow Y$  (Postulate 1), because  $P_{X \rightarrow Y}(x, y) = P(x)P(y)$ , and thus

$$D(P \| P_{X \rightarrow Y}) = D(P(X, Y) \| P(X)P(Y)) = I(X; Y).$$

**P2:** Note that the definition of  $\mathfrak{C}_{k \rightarrow l}$  refers to the impact of the cut on the *entire* joint distribution  $P(X_1, \dots, X_n)$ . It is therefore not obvious that  $\mathfrak{C}$  in fact only depends on the joint distribution of  $X_l$  and its parents, as required by Postulate 2. The following result shows that this is the case; furthermore it writes causal strength as an expression that is convenient for practical applications discussed later.

**Lemma 4** (causal strength as local relative entropy). *Causal strength  $\mathfrak{C}_{k \rightarrow l}$  can be written as the following relative entropy distance or conditional relative entropy distance:*

$$\begin{aligned} \mathfrak{C}_{k \rightarrow l} &= D[P(X_l, PA_l) \| P_S(X_l, PA_l)] \\ &= \sum_{pa_l} D[P(X_l | pa_l) \| P_S(X_l | pa_l)] P(pa_l) = D[P(X_j | PA_j) \| P_S(X_j | PA_j)]. \end{aligned}$$

Note that  $P_S(X_l|pa_l)$  actually depends on the reduced set of parents  $PA_l \setminus X_k$  only, but it is more convenient for the notation and the proof to keep the formal dependence on all  $PA_l$ .

*Proof.* Due to

$$P(X_1, \dots, X_n) = \prod_j P(X_j|PA_j),$$

and

$$P_S(X_1, \dots, X_n) = \prod_j P_S(X_j|PA_j),$$

we have

$$D(P||P_S) = \sum_j D[P(X_j|PA_j) || P_S(X_j|PA_j)] .$$

For all  $j \neq l$  we have  $D[P(X_j|PA_j)||P_S(X_j|PA_j)] = 0$ , because  $P(X_l|PA_l)$  is the only conditional that is modified by the deletion.  $\square$

**P3:** Apart from demonstrating the postulated inequality, the following result shows that we have the equality  $\mathfrak{C}_{X \rightarrow Y} = I(X; Y | PA_Y^X)$  for independent causes. Moreover, it provides an information theoretic interpretation for the additional term that occurs for dependent causes. To keep notation simple, we have restricted our attention to the case where  $Y$  has only two causes  $X$  and  $Z$ , but  $Z$  can also be interpreted as representing all parents of  $Y$  other than  $X$ .

**Theorem 5** (decomposition of causal strength). *For the DAGs in Figure 2 we have*

$$\mathfrak{C}_{X \rightarrow Y} = I(X; Y | Z) + D[P(Y|Z) || P_{X \rightarrow Y}(Y|Z)] . \quad (14)$$

*If  $X$  and  $Z$  are independent, the second term vanishes.*

*Proof.* Due to  $P_{X \rightarrow Y}(x, y, z) = \sum_{x'} P(y|x', z)P(x')P(x, z)$  we have

$$\begin{aligned} D(P||P_{X \rightarrow Y}) &= \sum_{x, y, z} P(y|x, z)P(x|z)P(z) \log \frac{P(y|x, z)}{\sum_{x'} P(y|x', z)P(x')} \\ &= \sum_{x, y, z} P(y|x, z)P(x|z)P(z) \log \frac{P(y|x, z)}{P(y|z)} \\ &\quad + \sum_{x, y, z} P(y|x, z)P(x|z)P(z) \log \frac{P(y|z)}{\sum_{x'} P(y|x', z)P(x')} \\ &= I(X; Y | Z) + D[P(Y|Z) || P_{X \rightarrow Y}(Y|Z)] . \end{aligned}$$

To see that the second term vanishes for independent  $X, Z$ , we observe  $P_{X \rightarrow Y}(Y|Z) = P(Y|Z)$  because

$$P_{X \rightarrow Y}(y|z) = \sum_x P(y|x, z)P(x) = \sum_x P(y|x, z)P(x|z) = P(y|z) .$$

$\square$

Theorem 5 states that conditional mutual information underestimates causal strength. Assume, for instance, that  $X$  and  $Z$  are always equal because  $Z$  has such a strong influence on  $X$  that it is an exact copy of it. Then  $I(X; Y | Z) = 0$  because knowing  $Z$  leaves no uncertainty about  $X$ . To see that causal influence cannot always be zero just because  $X$  and  $Z$  coincide, we consider the limiting case where the influence of  $Z$  on  $Y$  gets weaker

and weaker, while keeping the strong influence on  $X$ . Then we obtain  $Z \rightarrow X \rightarrow Y$  as limiting DAG. We have already seen that  $\mathfrak{C}_{X \rightarrow Y} = I(X; Y)$  in this case. In other words, strong dependences between the causes  $X$  and  $Z$  makes the influence of cause  $X$  almost invisible when looking at the conditional mutual information  $I(X; Y | Z)$  only. The second term in (14) corrects for the underestimation. When  $X$  depends deterministically on  $Z$ , it is even the only remaining term.

To provide a further interpretation of Theorem 5, we recall that  $I(X; Y | Z)$  can be seen as the impact of ignoring the edge  $X \rightarrow Y$ , see remarks around eq. (9). Then the impact of cutting  $X \rightarrow Y$  is given by the impact of ignoring this link plus the impact the cutting has on the conditional  $P(Y | Z)$ . In the appendix we show that this result generalizes to cutting and ignoring multiple links.

Finally, we collect together some nice properties of causal influence in the following theorem:

**Theorem 6** (relation between strength of sets and single arrows).

*The causal influence given in Definition 2 satisfies additivity on targets, locality, and monotonicity.*

**a) Additivity regarding targets.**

*Given set of arrows  $S$ , let  $S_i = \{s \in S | \text{trg}(s) = X_i\}$ , then*

$$\mathfrak{C}_S(P) = \sum_i \mathfrak{C}_{S_i}(P).$$

**b) Locality.**

*Every  $\mathfrak{C}_{S_i}$  only depends on the conditional  $P(X_i | PA_i)$  and the joint distribution of all parents  $P(PA_i)$ .*

**c) Monotonicity.**

*Given sets of arrows  $S_1 \subset S_2$  targeting single node  $Z$ , such that the source nodes in  $S_2$  are jointly independent. Then we have*

$$\mathfrak{C}_{S_1}(P) \leq \mathfrak{C}_{S_2}(P).$$

*Proof.* Appendix A.2. □

The intuitive meaning of these properties is as follows. Part (a) says that causal influence is additive if the arrows have different targets. Otherwise, we can still decompose the set  $S$  into equivalence classes of arrows having the same target and obtain additivity regarding the decomposition. We will show in Subsection 4.4 that general additivity fails. Part (b) is an along of P2 for multiple arrows. According to (c), the strength of a subset of arrows cannot be smaller than the strength of its superset, provided that there are no dependences among the parent nodes.

## 4.4 Examples and paradoxes

Although Theorem 6 shows that causal influence behaves nicely in many situations, there remain some examples where the results are somewhat counterintuitive. We collect these here.

**Failure of subadditivity:** The strength of a set of arrows is not bounded from above by the sum of strength of the single arrows. It can even happen that removing one arrow from a set has no impact on the joint distribution while removing all of them has significant impact:

**Example 2** (error correcting code). *Let  $E$  and  $D$  be binary variables that we call “encoder” and “decoder” (see figure 6) communicating over a channel that consists of the bits*

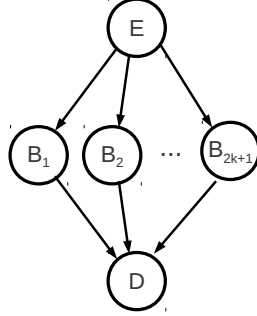


Figure 6: Causal structure of an error-correcting scheme: the encoder generates  $2k + 1$  bits from a single one. The decoder decodes the  $2k + 1$  bit words into a single bit again.

$B_1, \dots, B_{2k+1}$ . Using the simple repetition code, all  $B_j$  are just copies of  $E$ . Then  $D$  is set to the logical value that is attained by the majority of  $B_j$ . This way,  $k$  errors can be corrected, i.e., removing  $k$  or less of the links  $B_j \rightarrow E$  has no effect on the joint distribution, i.e.,  $P_S = P$  for  $S := (B_1 \rightarrow D, B_2 \rightarrow D, \dots, B_k \rightarrow D)$ , hence  $\mathfrak{C}_S(P) = 0$ . In words: removing  $k$  or less arrows is without impact, but removing all of them is, of course. After all, the arrows jointly generate the dependence  $I(E : D) = I(E : B_1, \dots, B_k, D) = 1$ .

Clearly, the outputs of  $E$  causally influence the behavior of  $D$ . We therefore need to consider interventions that destroy many arrows at once if we want to capture the fact that their joint influence is non-zero.

Thus, causal influence of arrows is *not* subadditive: the strength of each arrow  $E \rightarrow B_j$  is zero, but the strength of the set of all  $E \rightarrow B_j$  is 1 bit.

**Failure of superadditivity:** The following example reveals an opposing phenomenon, where the causal strength of a set is smaller than the sum of the single arrows:

**Example 3** (XOR-gate with COPY).

The causal influence of each arrow targeting an XOR-gate individually is the same as the causal influence of both arrows taken together:

$$\mathfrak{C}_{X \rightarrow Z}(P) = \mathfrak{C}_{Y \rightarrow Z}(P) = \mathfrak{C}_{X \rightarrow Z, Y \rightarrow Z}(P) = 1 \text{ bit.}$$

**Strong influence without dependence:** Revisiting our XOR-example is also instructive because it demonstrates an extreme case of confounding where  $I(X; Y | Z)$  vanishes but causal influence is strong.

**Example 4** (XOR-gate with copy).

Consider the DAG *a*) in figure 2 and let the relation between  $X, Y, Z$  be given by the structure equations

$$\begin{aligned} X &= Z, \\ Y &= X \oplus Z. \end{aligned}$$

Removing  $X \rightarrow Y$  yields

$$P_{X \rightarrow Y}(x, y, z) = P(x)P(y)P(z|x, y),$$

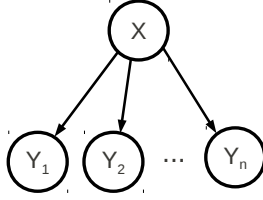


Figure 7: Broadcasting one bit from one node to multiple nodes.

where  $P(x) = P(y) = 1/2$  and  $P(z|x, y) = \delta_{z-x \oplus y}$ . It is easy to see that

$$D(P \parallel P_{X \rightarrow Y}) = 1,$$

because  $P$  is a uniform distribution over 2 possible triples  $(x, y, z)$ , whereas  $P_{X \rightarrow Y}$  is a uniform distribution over 4 combinations.

The impact of cutting the edge  $X \rightarrow Y$  is remarkable: both distributions, the observed one  $P$  as well as the post-cutting distribution  $P_S$ , factorize  $P_S(X, Y, Z) = P_S(X, Z)P_S(Y)$  and  $P(X, Y, Z) = P(X, Z)P(Y)$ . Cutting the edge keeps the product structure and only changes the marginal distribution of  $Y$ .

**Strong effect of little information:** The following example considers multiple arrows and shows that their joint strength may even be strong when they carry the same small amount of information:

**Example 5** (broadcasting).

Consider a single source  $X$  with many targets  $Y_1, \dots, Y_n$  such that each  $Y_i$  copies  $X$ , see Figure 7. Assume  $P(y_0 = 0) = P(y_0 = 1) = \frac{1}{2}$ . If  $S$  is the set of all arrows  $X \rightarrow Y_j$  then  $\mathfrak{C}_S = n$ . Thus, the single node  $X$  exerts  $n$  bits of causal influence on its dependents.

## 5 Causal influence between two time series

### 5.1 Definition

Since causal analysis of time series is of high practical importance, we devote a section to this case. For some fixed  $t$ , we introduce the short notation  $X \rightarrow Y_t$  for the set of all arrows that point to  $Y_t$  from some  $X_s$  with  $s < t$ . Then

$$\mathfrak{C}_{X \rightarrow Y_t}$$

measures the impact of deleting all these arrows. We propose to replace transfer entropy with this measure since it does not suffer from the drawbacks described in Subsection 3.2.

Subsection 4.2 describes how to estimate causal strength from finite data for one arrow and briefly mentions how this generalizes to set of arrows. To keep this section self-consistent, we briefly rephrase the description for the case of time series.

Suppose we have learned the structure equation model

$$Y_t = f_t(X_{t-1}, X_{t-2}, \dots, X_{t-p}, E_t),$$

from observed data  $(x_t, y_t)_{t \leq 0}$ , where the noise variables  $E_t$  are jointly independent and independent of  $X_t, X_{t-1}, \dots, Y_{t-1}, Y_{t-2}, \dots$ . Assume, moreover, that we have inferred the corresponding values  $(e_t)_{t \leq 0}$  of the noise.



To generate virtual observations via permutations we either need  $k$  samples for each time instant  $t$ , or the time series must be approximately stationary over a sufficiently large time interval. In the first case, we choose random permutations  $\pi_1, \dots, \pi_p \in S_k$  that permute the  $k$  samples within each time instant  $t$ . In the second case, we choose some number  $m \gg p$  and formally treat  $(x_{s-jm})_{j=1}^k$  as a sample, on which the permutations act. The permutations generate virtual observations

$$\tilde{x}_{t-1}, \dots, \tilde{x}_{t-p},$$

from which we compute the virtual  $Y_t$  values

$$\tilde{y}_t := f_t(\tilde{x}_{t-1}, \dots, \tilde{x}_{t-p}, e_t).$$

Then we estimate the relative entropy distance between the joint distribution of

$$Y_t, X_{t-1}, \dots, X_{t-p}$$

given by the real observations  $y_t, x_{t-1}, \dots, x_{t-p}$  and the virtual ones  $\tilde{y}_t, \tilde{x}_{t-1}, \dots, \tilde{x}_{t-p}$ .

## 5.2 Comparison of causal influence with transfer entropy

We first recall the example given by [10] showing a problem with transfer entropy (Subsection 3.2). Assume that the variables  $X_t, Y_t$  in figure 3, right, are binary and the transition from  $X_{t-1}$  to  $Y_t$  is a perfect copy and likewise the transition from  $Y_{t-1}$  to  $X_t$ . Assume, moreover, that the system has been initialized such that, with probability  $1/2$ , all variables are 1 and with probability  $1/2$  all are zero. Then the set  $X \rightarrow Y_t$  is the singleton  $S := \{X_{t-1} \rightarrow Y_t\}$ . Using Lemma 4, we have

$$\mathfrak{C}_{X_{t-1} \rightarrow Y_t} = D[P(Y_t, X_{t-1}) \| P_S(Y_t, X_{t-1})].$$

Since  $Y_t$  is a perfect copy of  $X_{t-1}$ , we have

$$P(y_t, x_{t-1}) = \begin{cases} 1/2 & \text{for } x_{t-1} = y_t \\ 0 & \text{otherwise} \end{cases}$$

into

$$P_S(y_t, x_{t-1}) = 1/4 \quad \text{for } (y_t, x_{t-1}) \in \{0, 1\}^2.$$

One easily checks  $D(P \| P_S) = 1$ .

Note that the example is somewhat unfair, since it is *impossible* to distinguish the structure equations from the case where  $X_{t+1}$  is the opposite of  $X_t$  and similarly for  $Y$ , no matter how many observations are performed. Thus, from observing the system it is impossible to tell whether or not  $X$  is exerting an influence on  $Y$ . However, the following modification shows that transfer entropy goes quantitatively still wrong if small errors are introduced:

**Example 6** (perturbed transfer entropy counterexample).

*Perturb Ay and Polani's example by having  $X_t$  copy  $Y_{t-1}$  correctly with probability  $p = 1 - \epsilon$ . Set node  $X_t$ 's transitions as Markov matrix*

$$\left( \begin{array}{c|cc} & x_t = 0 & x_t = 1 \\ \hline y_{t-1} = 0 & 1 - \epsilon & \epsilon \\ y_{t-1} = 1 & \epsilon & 1 - \epsilon \end{array} \right),$$

*and similarly for the transition from  $Y_{t-1}$  to  $X_t$ .*

*The transfer entropy from  $X$  to  $Y$  at time  $t =$  is*

$$TE := I(X_{(-\infty, t-1]}; Y_t | Y_{(-\infty, t-1]}) = I(X_{t-1}; Y_t | Y_{t-1}),$$

where we have used

$$Y_t \perp\!\!\!\perp X_{t-2}, X_{t-3}, Y_{t-1}, Y_{t-2}, \dots | X_{t-1}.$$

Some calculations show

$$TE = (1 - \epsilon)^2 \cdot \log \frac{1 - \epsilon}{1 - 2\epsilon + 2\epsilon^2} + \epsilon^2 \cdot \log \frac{\epsilon}{1 - 2\epsilon + 2\epsilon^2} + \epsilon(1 - \epsilon) \cdot \log \frac{1}{4\epsilon(1 - \epsilon)},$$

which tends to zero for  $\epsilon \rightarrow 0$ , in agreement with the unperturbed example. Causal influence, on the other hand, reads

$$\mathfrak{C} = (1 - \epsilon) \cdot \log(2 - 2\epsilon) + \epsilon \cdot \log(2\epsilon),$$

which tends to 1 for  $\epsilon \rightarrow 0$ . Hence, causal influence detects the causal interactions between  $X$  and  $Y$  based on empirical data, whereas transfer entropy is not. Thanks to the perturbation, the joint distribution tells us the kind of causal relations by which it is generated. For large enough samples, the strog discrepancy between transfer entropy and our causal strength thus becomes apparent.

## A Appendix: Further properties of causal strength

### A.1 Decomposition into conditional relative entropies

The following result generalizes Lemma 4 to the case where  $S$  contains more than one edge. It shows that causal strength, defined via the relative entropy between two distributions on the entire DAG, decomposes into a sum of conditional relative entropies, each referring to the conditional distribution of one of the target nodes given its parents:

**Lemma 7** (causal influence decomposes into a sum of expectations).

The causal influence of a set of arrows  $S$  can be rewritten

$$\mathfrak{C}_S(P) = \sum_{j \in \text{trg}(S)} D \left( P(X_j | PA_j) \left\| \sum_{pa_j^S} P(X_j | PA_j^{\bar{S}}, pa_j^S) \cdot P_{\Pi}(pa_j^S) \right. \right),$$

where  $\text{trg}(S)$  denotes the target nodes of arrows in  $S$ .

An intuitive implication of this result is Theorem 6 in the main text, whose proof is given in the next section.

*Proof.* By definition,  $\mathfrak{C}_S(P) = D(P \| P_S)$  and we obtain

$$D(P \| P_S) = D \left( \prod_j P(X_j | PA_j) \left\| \prod_{j=1}^n \sum_{pa_j^S} P(X_j | PA_j^{\bar{S}}, pa_j^S) \cdot P_{\Pi}(pa_j^S) \right. \right) = \quad (15)$$

$$D \left( \prod_{j \in \text{trg}(S)} P(X_j | PA_j) \left\| \prod_{j \in \text{trg}(S)} \sum_{pa_j^S} P(X_j | PA_j^{\bar{S}}, pa_j^S) \cdot P_{\Pi}(pa_j^S) \right. \right) = \quad (16)$$

$$\sum_{j \in \text{trg}(S)} D \left( P(X_j | PA_j) \left\| \sum_{pa_j^S} P(X_j | PA_j^{\bar{S}}, pa_j^S) \cdot P_{\Pi}(pa_j^S) \right. \right). \quad (17)$$

(15) = (16): Only the distributions of elements targeted by arrows in  $S$  are affected by the marginalization and therefore contribute to the causal influence; others play no role

and cancel out in the logarithms. Nodes in  $pa_S$  cancel out of the logarithm for the same reason. However, what remains inside the logarithm is a function of these values, hence the expectation over these nodes is nontrivial.

(16) = (17): Products go to sums by definition of relative entropy.  $\square$

## A.2 Proof of Theorem 6

*Proof.* Parts (a) and (b) follow from Lemma 7 since  $\mathfrak{C}_{S_i}(P)$  is the  $i$ th summand in (17), which obviously depends on  $P(X_i|PA_i)$  and  $P(PA_i)$  only. To prove part (c), start with the special case where  $G = \{X \rightarrow Z, Y \rightarrow Z\}$  and the sets are  $S_1 = \{Y \rightarrow Z\}$  and  $S_2 = \{X \rightarrow Z, Y \rightarrow Z\}$ :

$$\begin{aligned} \mathfrak{C}_{S_2}(P) &= \sum_{x,y} P(x)P(y)D \left( P(Z|x,y) \left\| \sum_{x,y} P(Z|x,y)P(x)P(y) \right\| \right) \\ &= \sum_{x,y} P(x)P(y)D \left( P(Z|x,y) \left\| \sum_y P(Z|x,y)P(y) \right\| \right) \\ &\quad + \sum_{x,y,z} P(x)P(y)P(z|x,y) \log \frac{\sum_y P(z|x,y)P(y)}{\sum_{x,y} P(z|x,y)P(x)P(y)} \\ &= \mathfrak{C}_{S_1}(P) + \sum_x P(x)D \left( \sum_y P(Z|x,y)P(y) \left\| \sum_{x',y'} P(Z|x',y')P(x')P(y') \right\| \right) \end{aligned}$$

In this case the proposition follows since  $D(R\|Q) \geq 0$  for any  $R$  and  $Q$ .

In the general case, the independence of the parents of  $Z$  implies

$$P(z, pa_Z) = P(z|pa_Z) \cdot P_{\Pi}(pa_Z^{S_2}) \cdot P(pa_Z^{\bar{S}_2}|pa_Z^{S_2}).$$

It follows that

$$\begin{aligned} \mathfrak{C}_{S_2}(P) &= \sum P_{\Pi}(pa_Z^{S_2}) \cdot P(pa_Z^{\bar{S}_2}|pa_Z^{S_2}) \cdot D \left( P(Z|pa_Z) \left\| P_{S_2}(Z|pa_Z^{\bar{S}_2}) \right\| \right) \\ &= \sum P_{\Pi}(pa_Z^{S_2}) \cdot P(pa_Z^{\bar{S}_2}|pa_Z^{S_2}) \cdot D \left( P(Z|pa_Z) \left\| P_{S_1}(Z|pa_Z^{\bar{S}_1}) \right\| \right) \\ &\quad + \sum P_{\Pi}(pa_Z^{S_2}) \cdot P(pa_Z^{\bar{S}_2}|pa_Z^{S_2}) \cdot P(z|pa_Z) \cdot \log \frac{P_{S_1}(z|pa_Z^{\bar{S}_1})}{P_{S_2}(z|pa_Z^{\bar{S}_2})}. \end{aligned}$$

The coefficient of the logarithm,  $\sum P(z|pa_Z) \cdot P_{\Pi}(pa_Z^{S_2}) \cdot P(pa_Z^{\bar{S}_2}|pa_Z^{S_2})$ , can be written as a sum of terms of the form  $P_{S_1}(z|pa_Z^{\bar{S}_1}) = \sum_{pa_Z^{S_1}} P(z|pa_Z)P_{\Pi}(pa_Z^{S_1})$  since we have assumed that the sources are independent. Consequently,

$$\mathfrak{C}_{S_2}(P) = \mathfrak{C}_{S_1}(P) + \sum P_{\Pi}(pa_Z^{S_2 \setminus S_1}) \cdot P(pa_Z^{\bar{S}_2}|pa_Z^{S_2 \setminus S_1}) \cdot D \left( P_{S_1}(Z|pa_Z^{\bar{S}_1}) \left\| P_{S_2}(Z|pa_Z^{\bar{S}_2}) \right\| \right)$$

and the proposition follows because relative entropy is non-negative.  $\square$

## A.3 Causal influence measures controllability

Causal influence is intimately related to control. Suppose an experimenter wishes to understand interactions between components of a complex system. She is able to observe nodes  $Y$  and  $Z$ , and manipulate node  $X$ . To what extent can she control node  $Y$ ? The notion of control has been formalized information-theoretically in [14]:

**Definition 5** (perfect control).

Node  $Y$  is perfectly controllable by node  $X$  at  $Z = z$  if, given  $z$ ,

- i) states of  $Y$  are a deterministic function of states of  $X$ ; and
- ii) manipulating  $X$  gives rise to all states of  $Y$ .

Perfect control can be elegantly characterized:

**Theorem 8** (information-theoretic characterization of perfect controllability).

A node  $Y$  with inputs  $X$  and  $Z$  is perfectly controllable by  $X$  alone for  $Z = z$  iff there exists a Markov transition matrix  $R(x|z)$  such that

$$H(Y|z, do X) := \sum_x R(x|z) H(Y|z, do x) = 0, \text{ and} \quad (C1)$$

$$\sum_{x \in X} P(y|z, do x) R(x|z) \neq 0 \text{ for all } y. \quad (C2)$$

Here,  $H(Y|z, do x)$  denotes the conditional Shannon entropy of  $Y$ , given that  $Z = z$  has been observed and  $X$  has been set to  $x$ .

*Proof.* The theorem restates the criteria in the definition. For a proof, see [14].  $\square$

It is instructive to compare Theorem 8 to our measure of causal influence. The theorem highlights two fundamental properties of perfect control. First, (C1), perfect control requires there is no variation in  $Z$ 's behavior given the choice of  $y$ . Second, (C2), perfect control requires that all potential outputs of  $Z$  can be induced by manipulating node  $Y$ . This suggests a measure of the *degree* of control should reflect (i) the variability in  $Z$ 's behavior that cannot be eliminated by imposing  $Y$  values and (ii) the size of the repertoire of behaviors that can be induced on the target by manipulating a source.

If  $X$  and  $Y$  are independent then by Theorem 5

$$\mathfrak{C}_{X \rightarrow Y}(P) = I(X; Y|Z) = H(Y|Z) - H(Y|X, Z).$$

The first term,  $H(Y|Z)$ , quantifies size of the repertoire of outputs of  $Y$  averaged over inputs from  $Z$ . It corresponds to requirement (C2) in the characterization of perfect control: that  $\sum_x P(y|z, do x) R(x|z) > 0$  for all  $z$ . Specifically, the causal influence, interpreted as a measure of the degree of controllability, increases with the size of the (weighted) repertoire of outputs that can be induced by manipulations.

The second term,  $H(Y|X, Z)$  (which coincides with  $H(Y|Z, do X)$  here), quantifies the variability in  $Y$ 's behavior that cannot be eliminated by controlling  $X$ . It corresponds to requirement (C1) in the characterization of perfect control: that remaining variability should be zero. Causal influence increases as the variability  $H(Y|Z, do X) = \sum_z P(z) H(Y|z, do X)$  tends towards zero.

#### A.4 Causal strength majorizes observed dependence

Recalling that  $P(X_1, \dots, X_n)$  factorizes into  $\prod_j P(X_j|PA_j)$  with respect to the true causal DAG  $G$ , one may ask how much error would arise if one was not aware of all causal influences and erroneously worked with a DAG where interactions across some set of arrows  $S$  in the true DAG  $G$  are hidden. The conditionals with respect to the reduced set of parents define a different joint distribution:

**Definition 6** (partially observed distribution).

Given distribution  $P$ , Markovian with respect to  $G$ , and set of arrows  $S$ , let the partially observed distribution (where interactions across  $S$  are hidden) for node  $X_j$  be

$$\tilde{P}_S(x_j|pa_j^{\bar{S}}) = \sum_{pa_j^S} P(x_j|pa_j^S, pa_j^{\bar{S}}) P(pa_j^S|pa_j^{\bar{S}}).$$

Let the partially observed distribution for all the nodes be the product

$$\tilde{P}_S(x_1, \dots, x_n) = \prod_j \tilde{P}_S(x_j | pa_j^{\bar{S}}).$$

*Remark 4.* Intuitively, the observed influence of a set of arrows should be quantified by comparing the data available to an observer who can see the entire DAG with the data available to an observer who sees all the nodes of the graph, but only some of the arrows. Definition 6 formalizes “seeing only some of the arrows”.

The definition of the *observed dependence* of a set of arrows takes the same general form as for causal influence. However, instead of inserting noise on the arrows, we instead simply prevent ourselves from seeing them:

**Definition 7** (observed influence).

Given distribution  $P$  Markovian with respect to  $G$  and set of arrows  $S$ , let the observed influence of the arrows in  $S$  be

$$\mathfrak{D}_S(P) := D(P \| \tilde{P}_S).$$

The following result generalizes Theorem 5:

**Theorem 9** (causal influence majorizes observed dependence).

Causal influence decomposes into observed influence plus a non-negative term quantifying the divergence between the partially observed and interventional distributions

$$\mathfrak{C}_S(P) = \mathfrak{D}_S(P) + D(\tilde{P}_S \| P_S).$$

The theorem shows that “snapping upstream dependencies” by using purely local data – i.e. by marginalizing using the distribution of the source node  $P(X_i)$  rather than the conditional  $P(X_i | PA_i)$  – is essential to quantifying causal influence.

*Proof.* Expand  $\mathfrak{C}_S(P)$  as

$$\begin{aligned} D(P \| P_S) &= \sum P(x_1 \dots x_n) \log \frac{P(x_1 \dots x_n)}{P_S(x_1 \dots x_n)} \\ &= \sum P(x_1 \dots x_n) \log \frac{P(x_1 \dots x_n)}{\tilde{P}_S(x_1 \dots x_n)} + \sum P(x_1 \dots x_n) \log \frac{\tilde{P}_S(x_1 \dots x_n)}{P_S(x_1 \dots x_n)}. \end{aligned}$$

By the proof of Lemma 7, the second term can be written as

$$\begin{aligned} &\sum_{j \in \text{trg}(S)} D(\tilde{P}_S(X_j | PA_j^{\bar{S}}) \| P_S(X_j | PA_j^{\bar{S}})) \\ &= \sum_{j \in \text{trg}(S)} D(\tilde{P}_S(X_j | PA_j^{\bar{S}}) \| P_S(X_j | PA_j^{\bar{S}})) = D(\tilde{P}_S \| P_S), \end{aligned}$$

where the last equality also follows by the proof of Lemma 7.

Causal influence is thus observed influence plus a correction term that quantifies the divergence between the partially observed and interventional distributions. The correction term is non-negative since it is a Kullback-Leibler divergence.  $\square$

## B Another option to define causal strength

We now discuss a slightly different approach to defining the strength of an arrow. Although it has many nice properties and is quite intuitive, it fails satisfying Postulate 3. To define

the strength of the arrow  $X \rightarrow Y$  we consider  $X$  and its parents  $PA_Y$  and define a modified joint distribution on  $PA_Y$  by

$$P'(X, PA_Y^X) := P(X)P(PA_Y^X)P(Y|PA_Y^X).$$

In words: we remove the dependences between  $X$  and the other parents of  $PA_Y^X$ . Then we define causal strength by the conditional mutual information

$$I_{P'}(X; Y | PA_Y^X)$$

with respect to the modified distribution. The modification can be thought of describing the post-interventional distribution where  $X$  is set to the values  $x$  according to the observed marginal distribution  $P(X)$ . To show that Postulate 3 is violated, we consider the case where two dependent variables  $X$  and  $Z$  influence  $Y$ . Let  $X$  consist of  $k + 1$  bits,  $Y$  be  $k$  bits and  $Z$  be just one bit. Call the first  $k$  bits of  $X$  the message and the remaining one the control bit. Define  $P(Y|X, Z)$  such that the message bits are copied to  $Y$  whenever both the control bit of  $X$  and the variable  $Z$  are set to 1. Otherwise, they are uniformly distributed on  $\{0, 1\}^k$ . To specify  $P(X, Z)$ , we first define a distribution on  $P(X_1, Z)$  by

$$P(x_1, z) = \begin{cases} 1/2 & \text{for } x_1 = z \\ 0 & \text{otherwise} \end{cases}$$

Then we set

$$P(X, Z) := P(X_1, Z)P(X_2, \dots, X_k),$$

where  $P(X_2, \dots, X_k)$  is the uniform distribution on  $\{0, 1\}^k$ . It is easy to see that

$$I(X; Y | Z) = k/2$$

because the  $k$  message bits are copied whenever  $Z = X_1 = 1$ , which happens with probability  $1/2$ . However, modifying  $P$  to  $P'$  breaks the coupling between  $X_1$  and  $Z$ , and  $X_1 = Z = 1$  only happens with probability  $1/4$ . Thus, the message bits are only copied with probability  $1/4$  and therefore

$$I_{P'}(X; Y | Z) = k/4 < I(X; Y | Z),$$

while Postulate 3 requires

$$I_{P'}(X; Y | Z) \geq I(X; Y | Z).$$

## C The problem of defining total influence

If  $X$  influences  $Y$  via directed paths other than a direct arrow, we may want to measure the total influence of  $X$  on  $Y$ .

However, we cannot quantify total influence by quantifying the impact of removing all the arrows on the directed paths connecting  $X$  and  $Y$ . To see this, consider the causal chain  $X \rightarrow Z \rightarrow Y$  and assume that  $X$  and  $Z$  are strongly coupled (e.g. by a copy operation), but  $Z$  and  $Y$  are weakly coupled (e.g. a very noisy copy operation). Then, removing both arrows  $X \rightarrow Z$  and  $Z \rightarrow Y$  has a large impact on  $P(X, Z, Y)$  although  $Y$  obtains almost no signal from  $X$ . In this simple case, total influence may be defined by replacing the path  $X \rightarrow Z \rightarrow Y$  with a single arrow and computing the direct influence in  $X \rightarrow Y$  after marginalizing  $P(X, Z, Y)$  to  $P(X, Y)$ . However, we do not see how to construct a general rule for shrinking total influence to a direct arrow. To describe the problem, consider DAG e) in figure 2. Since  $X$  influences  $Y$  only directly, we may want to consider  $\mathfrak{C}_{X \rightarrow Y} = I(X; Y | Z)$  also as the strength of the total influence. On the other hand, for case f), we would tend to consider  $I(X; Y)$  as the total influence. This is because

shrinking the DAG to a direct influence would represent the net effect of both influences, the direct one  $X \rightarrow Y$ , and the indirect one  $X \rightarrow Z \rightarrow Y$  into a single arrow  $X \rightarrow Y$ . The distribution  $P(X, Y)$  is simply given by marginalizing  $P(X, Z, Y)$ . On the other hand, DAG e) can be considered as a special instance of case f) by adding an irrelevant arrow to e). Inserting the “virtual edge” then changes the total influence from  $I(X; Y | Z)$  to  $I(X; Y)$ . If, for instance  $Y = X \oplus Z$  and  $X$  and  $Z$  are unbiased coins,  $I(X; Y) = 0$  but  $I(X; Y | Z) = 1$ .

## References

- [1] J. Pearl. *Causality*. Cambridge University Press, 2000.
- [2] S. L. Lauritzen. *Graphical models*. Oxford Statistical Science Series. Oxford University Press, Oxford, 1996.
- [3] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Lecture Notes in Statistics. Springer, New York, 1993.
- [4] P.W. Holland. Causal inference, path analysis, and recursive structural equations models. In C. Clogg, editor, *Sociological Methodology*, pages 449–484. American Sociological Association, Washington DC, 1988.
- [5] R. Northcott. Can ANOVA measure causal strength? *The Quarterly Review of Biology*, 83(1):47–55, 2008.
- [6] R.C. Lewontin. Annotation: the analysis of variance and the analysis of causes. *American Journal Human Genetics*, 26(3):400–411, 1974.
- [7] T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Telecommunications, New York, 1991.
- [8] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, July 1969.
- [9] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.
- [10] N. Ay and D. Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(1):17–41, 2008.
- [11] J. Lizier and M. Prokopenko. Differentiating information transfer and causal effect. *The European Physical Journal B*, 73:605–615, 2010.
- [12] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Proceedings of the conference Neural Information Processing Systems (NIPS) 2008*, Vancouver, Canada, 2009. MIT Press. [http://books.nips.cc/papers/files/nips21/NIPS2008\\_0266.pdf](http://books.nips.cc/papers/files/nips21/NIPS2008_0266.pdf).
- [13] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. <http://uai.sis.pitt.edu/papers/11/p589-peters.pdf>.
- [14] H. Touchette and S. Lloyd. Information-theoretic approach to the study of control systems. *Physica A*, 331:140–172, 2004.